Abstract Anaphors in German and English

Stefanie Dipper[1], Christine Rieger[2], Melanie Seiss[2], and
Heike Zinsmeister[2]

[1] Ruhr-University Bochum, 44780 Bochum, Germany
[2] University of Konstanz, 78457 Konstanz, Germany

Abstract. Abstract anaphors refer to abstract referents such as facts or events. Automatic resolution of this kind of anaphora still poses a problem for language processing systems. The present paper presents a corpus-based comparative study on German and English abstract anaphors and their antecedents to gain further insights into the linguistic properties of different anaphor types and their distributions. To this end, parallel texts from the Europarl corpus have been annotated with functional and morpho-syntactic information. We outline the annotation process and show how we start out with a small set of well-defined markables in German. We successively expand this set in a cross-linguistic bootstrapping approach by collecting translation equivalents from English and using them to track down further forms of German anaphors, and, in the next turn, in English, etc.

Keywords: Abstract anaphora, corpus annotation, contrastive linguistics

1 Introduction

Abstract anaphora denote anaphoric relations between some anaphoric expression and an antecedent that refers to an abstract object like an event or a fact.[1] The antecedents are normally expressed by verbal or clausal constructions, and sometimes also by their corresponding nominalizations. In the classical example by Byron [4], the pronoun *it* (underlined in (1a)) refers to an *event*: the migration of penguins to Fiji. In the alternative sequence, (1b), the demonstrative pronoun *that* refers to the *fact* that penguins migrate to Fiji in the fall.

(1) a. Each Fall, penguins migrate to Fiji. <u>It</u> happens just before the eggs hatch.
b. Each Fall, penguins migrate to Fiji. <u>That</u>'s why I'm going there next month.

Abstract anaphora are analyzed as discourse deixis within a mental discourse model [22, 24]. According to this approach, discourse units correspond to underspecified abstract referents which can be coerced into different types of referents

[1] We would like to thank the reviewers for helpful comments. Heike Zinsmeister's research was financed by Europäischer Sozialfonds in Baden-Württemberg.

when they are referred to in the text [2]. Abstract anaphora contribute to the coherence of a text in that they make previously mentioned events (or facts etc.) available for further modification in subsequent sentences. Compared to other tasks in Natural Language Processing such as tagging or parsing, automatic resolution of concrete anaphora is still a difficult challenge for language processing systems. Consequently, automatic resolution of abstract anaphora is an even harder task.

We pursue a corpus-based approach to investigate the properties that characterize different instantiations of abstract anaphora. In the long run, we envisage to derive features from the corpus annotation that will serve us to tackle the automatic resolution of abstract anaphors. In this paper we investigate what kind of anaphoric elements are employed to refer to abstract objects. The range of possible realizations includes pronouns, lexical NPs (e.g. *this issue, this situation*, etc.) and adverbials (e.g. *likewise*). We take a cross-linguistic, bootstrapping approach and present a comparative corpus study on the realization of abstract anaphora in a parallel corpus of English and German. We present results on the following question: to what extent do English and German use the same kind of strategies to refer to abstract objects.

The paper is organized as follows. In Sec. 2, we present related work. Sec. 3 provides a description of our approach: the corpus, methodological considerations, and the annotation procedure. Sec. 4 presents the results from our comparative study in detail while Sec. 5 discusses the results more generally. Sec. 6 concludes with an outline of future research.

2 Related Work

In comparison to work on nominal anaphora, considerably less research has focused on abstract anaphora. A recent overview of projects annotating abstract anaphora is provided by [8]. Studies based on English (monolingual) corpora, e.g., include [23, 16, 5, 10, 13, 18, 17]. Languages other than English have been studied by [12, 15] (Czech), [1] (Basque), and [9] (German).

Contrastive analyses based on multilingual *comparable* corpora have been made, e.g., by [19] for Spanish and Catalan, which investigates all kinds of pronouns and full NPs. The data shows that in Catalan, demonstrative pronouns are used slightly more frequently than personal pronouns to refer to abstract entities (thus reflecting tendencies that can be found also in English). In contrast, Catalan uses personal pronouns twice as much as demonstratives.

In a diachronic study of English data from the 17th–20th centuries, [3] finds that the use of the personal pronoun *it* as an abstract anaphor has decreased over time, and the demonstrative pronoun *that* came into use instead; throughout the entire period, *this* is rarely used as an abstract anaphor.

Annotation of *parallel* texts has been performed, e.g., by [21], who extract a French-Portuguese subcorpus from the parallel MLCC corpus. The MLCC corpus contains written questions asked by members of the European Parliament and the corresponding answers from the European Commission. [21] investigate the use of demonstrative NPs. Although French has a higher number of demonstratives, the overall results are highly similar, and French and Portuguese seem to share relevant syntactic and semantic properties.

[14] annotates pronominal abstract anaphora in Andersen's fairy tales in Danish (the original language), and their English and Italian translations. The data shows that whereas English mostly uses demonstrative pronouns to refer to abstract entities, there is no such preference in Danish and Italian, which also use personal pronouns quite often. In original Italian data, abstract anaphors occur less frequently than in the translations.

Our project deals with the annotation of the full range of abstract anaphora (including full NPs anaphors and anaphoric adverbs) in a parallel corpus in German and English. In this paper, we present the first two annotation rounds of a bi-directional bootstrapping approach which concentrates mainly on pronominal anaphors.

3 Our Study

3.1 The Corpus

For our study, we extracted about 100 German and English turns (contributions by German and English speakers) along with their sentence-aligned translations from the Europarl Corpus (Release v3, 1996–2006, [11]). The Europarl corpus consists of transcripts of European Parliament debates. Individual contributions ('turns') in the debates were delivered (and transcribed) in one of the official EU languages. Professional translators provided official EU translations.

The original contributions were spoken but might have been based on written scripts. Speakers had the option to edit the transcripts before publication. Hence, the register of the turns is of a mixed character, between spoken and a more standardized written language.

Preprocessing of the data included the addition of missing tags to indicate the speaker's original language. More importantly, it included tokenizing, POS tagging and chunking based on the TreeTagger [20].

We created two parallel subcorpora: (i) "DE-EN" based on German original turns and their aligned English translations; (ii) "EN-DE" based on English originals and German translations. DE-EN contains 94 German turns, with an average of 19.5 sentences per turn. The turns contain contributions by 61 German and Austrian speakers. The turns were randomly sampled from those turns of the German Europarl corpus that contain at least one markable, i.e. one of pronominal *dies, das, es* 'this, that, it' (see below). For the annotation task, all 871 markables in the turns were highlighted; among them, 223 were identified as abstract anaphors by the annotators (Ø 2.37 abstract anaphors per turn). 203 of them could be aligned with English equivalents.[2]

[2] The alignment is not complete since it is based on the automatic sentence alignment provided by Europarl, release v3, which does not contain alignments for all turns. If a translation is not literal, the turn structures of the parallel texts are not necessarily isomorphic.

EN-DE is about the same size as DE-EN. It contains 95 English turns with an average length of 21.0 sentences. 296 abstract anaphora were identified on the basis of 1,224 markables and aligned with their German translations (Ø 3.12 anaphors per turn).[3]

The results presented in Sec. 4 are based on the set of aligned anaphora pairs of both translation directions.

3.2 Methodological Considerations

One way to learn about the distribution of abstract anaphora would be to go through a text and check sentence by sentence whether it contains a reference to an abstract referent. We do not pursue this approach. Instead, we start out with a well-defined set of markables in the original language and collect all variants of translations on the side of the "target" language (the translation of the original language).

In the first round of annotation, we chose original texts from German, because in German —in contrast to English— one pronoun is unambiguously used as an abstract anaphor: the uninflected singular demonstrative pronoun *dies* ('this'). In addition to this, we defined as markables the (ambiguous) demonstrative pronoun *das* ('that') and the (ambiguous) third person neuter pronoun *es* ('it'). The target language was English.

For the second round of annotation, we considered the reversed translation direction: English original texts and their German translations. We extended our set of markables and included the adverbs *as, so* and *likewise*, because these adverbs frequently served as translations of German anaphors in the first round. We will apply this method of bootstrapping back and forth to extend the set of markables iteratively. For instance, in the third round, German pronominal adverbs (e.g. *davon* 'thereof') and the adverb *wie* ('as') will be added to the set of markables. In contrast to the first approach described above, this bootstrapping approach allows for a fast and efficient way of extracting anaphors in both languages.

3.3 Annotation Procedure

For cross-lingual annotation of German and English texts, two MMAX2 annotation windows were used, which were put side by side on the screen.[4]

The annotators were first asked to annotate the German text. For each anaphor, they had to specify its type (demonstrative or personal pronoun), function (subject, object, other) and position (pre-field, matrix, embedded, other).

Next the annotators checked whether some item could be identified in the corresponding English align unit which served a similar function as the German

[3] The slightly higher density of abstract anaphors in English is due to the fact that we extracted turns containing at least two markables and started out with the extended set of markables avaible after the first annotation round, see Sec. 3.2.

[4] MMAX2: `http://mmax2.sourceforge.net/`

anaphor. If such an item was found on the English side, it was marked and, similarly to the annotation of German, its type, function, and position were specified. For the annotation of English, the feature 'type' could be specified as: pronoun, NP, *likewise, so, as* or other); the features 'function' and 'position' have the same values as in German, except for the position *pre-field* ('Vorfeld'), which is replaced by a *topic* position in English. The English anaphoric item was linked to the German anaphor via the token-ID of the German anaphor.[5] Obviously, in the first round of annotation, only anaphors of a very restricted, predefined set were annotated, and only anaphors that were present in the German texts were considered at all. To complete the picture, we therefore looked at original English texts, too, and started out from English anaphors (as defined above) and searched for corresponding items in the German translations. This way, we came across new forms of abstract anaphors in German, which can be used in the bootstrapping approach, to search, again, for new forms in English.

4 Results

We start this section by testing two hypotheses: that English in general avoids the use of pronominal abstract anaphors, and that English prefers demonstrative pronouns to personal pronouns in abstract anaphora. We then compare the grammatical functions and positions of abstract anaphors in German and English.

4.1 Avoidance of Pronominal Abstract Anaphors in English

We used our annotations to test the hypothesis that English avoids the use of *pronominal* abstract anaphors. The results from the German-to-English ('DE-EN') and English-to-German ('EN-DE') annotations do not to support this hypothesis. Table 1 shows that in both directions, the majority of pronouns (65% and 70%) are translated to a pronoun in the target language, while a small part is translated to full NPs and the rest to some other expression (e.g. anaphoric adverbials).[6] The differences between the two translation directions are not statistically significant.

The data shows that both languages use pronominal abstract anaphors to a similar extent, but the uses overlap in around 70% of the cases only. One possible explanation could be that the contexts of the abstract anaphors are at the root of the discrepancies: while the contexts are semantically more or less equivalent (because one is the translation of the other), they can differ at the syntactic level, with the effect of disallowing a source pronoun in the target language.

[5] To ensure reliable annotations, annotation guidelines were provided, a detailed one for monolingual annotation, which includes tests for antecedents etc. [7], and more general guidelines, describing the process of bilingual annotation in two MMAX2 windows [6]. Due to space limitations, we cannot go into the details of the guidelines here.

[6] EN-DE: 39 pronominal adverbs are counted among the German pronouns.

Table 1. Translations of pronouns

	Pronoun-to-pronoun	Pronoun-to-other	Sum
DE-EN	65% (132)	35% (19 NPs, 52 other)	100% (203)
EN-DE	70% (173)	29% (18 NPs, 55 other)	100% (246)

We observe the following main differences in the translations of pronominal abstract anaphors from German to English and vice versa.

– there is no corresponding material in the translation, e.g. a different argument frame is employed, see Ex. (2)[7]
– use of full NPs rather than pronouns (*all these things, the whole thing, this approach, these measures, this situation, this thread* ...), see Ex. (3)
– use of adverbials or conjunctions (*likewise, so, as*), see Ex. (4)

(2) a. DE_o: Wenn dies nicht geschieht, verlieren wir das Vertrauen der Bürger.
EN_t: If we do not, the public will lose confidence in us.
DE-LIT: ... If this does not happen, the public will lose confidence in us.
b. EN_o: There are absolute assurances of that and provisions made for it in the White Paper.
DE_t: Hierfür sind absolute Sicherungsmaßnahmen vorgesehen, und das Weißbuch enthält die notwendigen Vorkehrungen.
DE-LIT: ... the White Paper lists the necessary provisions.

(3) a. DE_o: Das konnte durch die glänzende Vorsitzführung von Frau Cederschiöld, aber auch durch die sehr substanzielle Hilfe der Kommission abgewendet werden, und deswegen können wir diesem Kompromissergebnis zustimmen.
EN_t: Thanks to Mrs Cederschiöld's inspired leadership, but also due to the very substantial support from the Commission, this threat has been averted, so we can now vote in favour of this compromise result.
DE-LIT: ... this could be averted
b. EN_o: I do not necessarily support this.
DE_t: Diesem Standpunkt schließe ich mich nicht notwendigerweise an.
DE-LIT: This position I do not necessarily follow.

(4) a. DE_o: ... — auch das wurde bereits gesagt — ...
EN_t: As has also been said already, ...
DE-LIT: — this too has been said already —

[7] In the examples, the a.-examples stem from the DE-EN corpus, the b.-examples from the EN-DE corpus. The lines displayed first contain the original version, additionally marked by the subscript "o". The second lines, with subscript "t", show the corresponding translation from the Europarl corpus. The "DE-LIT" lines provide a literal translation of (parts of) the German lines.

b. EN_o: Whatever European Union policies flow from this conference at The Hague will have to come to this Parliament for debate, amendment and agreement, that is the European policies.
DE_t: So sieht es das europäische Regelwerk vor.
DE-LIT: So it is regulated by the European regulations.

4.2 Preference of Demonstrative Pronouns in English

Following [16, 14], we hypothesized that English prefers demonstrative pronouns to personal pronouns in abstract anaphora in comparison to other languages.

Fig. 1 shows the translation equivalents of pronoun types from both translation directions. The EN-DE bar plot indeed confirms that English prefers demonstrative pronouns ($> 80\%$).[8] The DE-EN bar plot, however, shows that German shows a similar preference. Such strong preferences did not show up for the languages studied by [14, 19] (Danish, Italian, Spanish, Catalan). In both directions, only about 2/3 of the demonstratives (DE-EN: 60%, EN-DE: 65%) are translated as such, and considerably less of the personal pronouns.

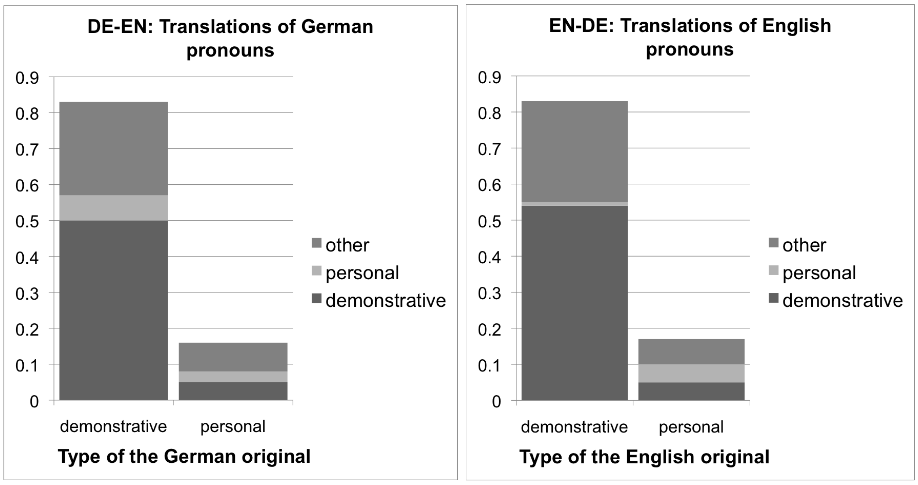

Fig. 1. Translation equivalents of the pronominal types (percentages). The columns encode the types of the original texts, the colors the types of the translated texts.

It is assumed that the personal pronoun *it* in English can only refer to events and states, but not, e.g., to situations or facts, see, e.g., [10]. This constraint does not seem to apply to German anaphors, which could explain part of the variance

[8] EN-DE: German pronominal adverbs are not considered here since their pronominal part is ambiguous between personal and demonstrative.

observed above. To validate such a hypothesis, we would need to annotate the semantic types of the abstract objects that are referred to by the anaphors.

Ex. (5a) shows a relevant type mismatch: German *es* ‘it’ refers to the *fact* that the states have not transposed the directive. The English translation uses the demonstrative *this* instead. A converse example is Ex. (5b): English *it* is translated by a German demonstrative.[9]

(5) a. DE_o: Frau Kommissarin, Sie haben jede Unterstützung dieses Parlaments, die Staaten, die diese Richtlinie nicht ordentlich umgesetzt haben, vor den EuGH zu bringen, es öffentlich zu machen und so den Druck dafür zu erzeugen, dass diese Richtlinie endlich umgesetzt wird.
EN_t: If, Commissioner, you want to bring before the ECJ those states that have not properly transposed this directive, in order to bring this out into the open and thus to bring pressure to bear in order to get this directive transposed at last, then this House is behind you all the way.
DE-LIT: . . . to bring it out into the open . . .

b. EN_o: The fact that an agreement was reached on very difficult issues should not be underestimated. It was a huge task.
DE_t: Die Tatsache, dass zu sehr schwierigen Fragen Übereinstimmung erzielt wurde, sollte nicht unterschätzt werden. Das war eine gigantische Aufgabe.
DE-LIT: . . . This was a huge task.

Comparing the uses of personal and demonstrative pronouns in English and German is hindered by the fact that the German neuter pronoun *es* ‘it’ is usually not used after prepositions and, instead, pronominal adverbs, such as *davon* ‘thereof’ or *daraus* ‘out of it’, are used—this holds for both concrete and abstract *es*-anaphors, see Ex. (6). Pronominal adverbs do not allow us to distinguish between personal or demonstrative use. Conversly, English seems to prefer personal to demonstrative pronouns after prepositions: *out of it/*that* [16].

(6) EN_o: The role of this Parliament is to ensure that the rules are complied with. That is what we should concentrate on.
DE_t: Die Aufgabe des Parlaments besteht darin, dafür zu sorgen, daß die Regeln eingehalten werden. Und genau darauf sollten wir uns konzentrieren.
DE-LIT: . . . And exactly thereon we should concentrate.

4.3 Function

In both languages, abstract anaphors (of the types that we have annotated up to now) most often occur in the subject position ($\geq 60\%$), see Fig. 2. The majority of subjects remain subjects (about 2/3), whereas only half of the objects are translated as such, in both directions. The overall picture of both translation directions is highly similar.

[9] It is not entirely clear to us to which kind of abstract object the anaphors refer to in Ex. (5).

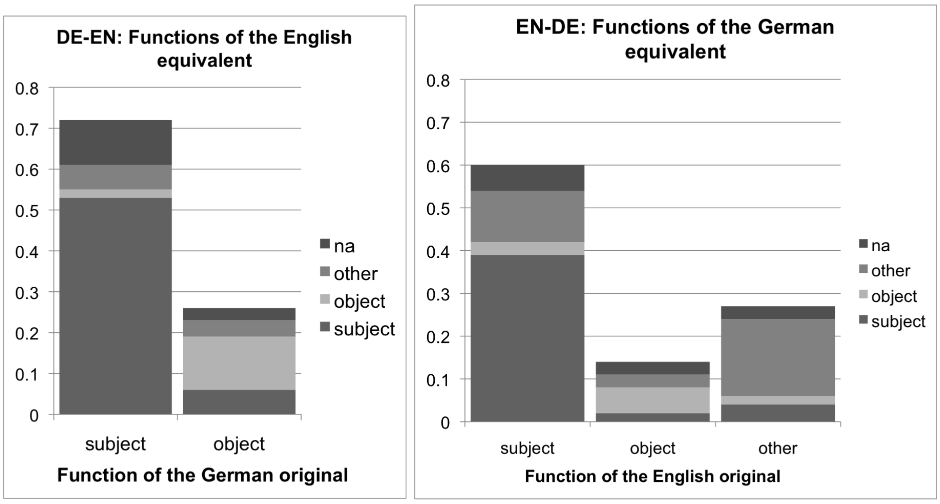

Fig. 2. Translation equivalents of functions (percentages)

4.4 Position

In the annotation, we distinguished between three different positions: the pre-field/topic position, a position within the matrix clause, and a position within the embedded clause. In Ex. (7), the original German anaphor is in an unmarked, post-verbal position. In contrast, its English counterpart has been realized in the marked topic position.

(7) DE_o: Man glaubte in verschiedenen europäischen Staaten, man müsste rasch handeln, man müsste die Amerikaner unterstützen. Ich verstehe das auch. Nur jetzt müssen wir wieder zur Rechtsstaatlichkeit zurückfinden ...
EN_t: It was believed in various European states that rapid action was called for and that we had to support the Americans, and that I can understand. Now, though, we have to get back to the rule of law ...
DE-LIT: ... and I understand that well. ...

According to Fig. 3, most abstract anaphors do not occur in embedded position. The figure further shows that the German pre-field position has other properties than the English topic position: The majority of German pre-field anaphors are translated as an ordinary matrix constituent in English. Conversely, English topicalized anaphors are usually translated to German pre-field anaphors. Columns 2 and 3 indicate that a minority of anaphors switch their position from a matrix clause into an embedded one, or vice versa.

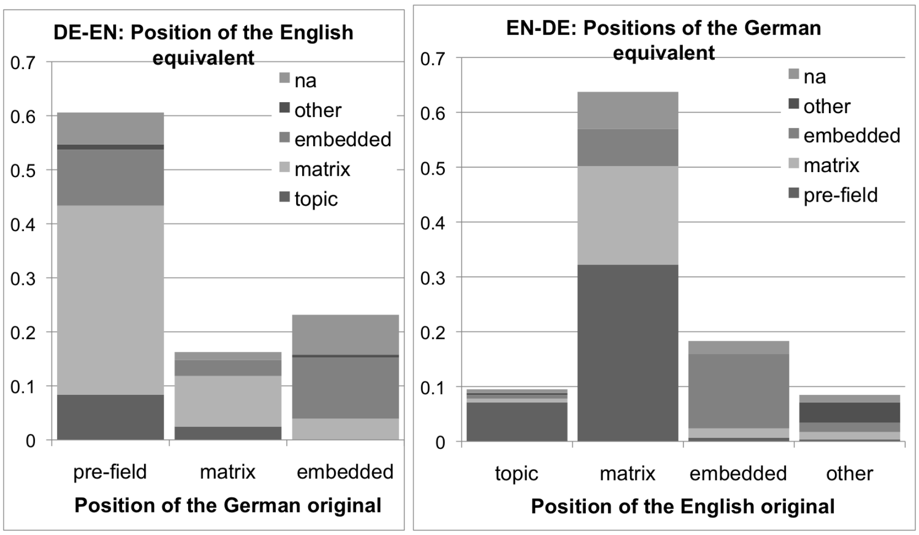

Fig. 3. Translation equivalents of positions (percentages)

5 Discussion

We performed a bidirectional comparison of the expression of abstract anaphora to interpret observed divergences between English and German.

The major finding of our study is that English and German pattern very much alike in contrast to findings on different language pairs. Despite the observed similarities, there are language-specific preferences that manifest themselves in cross-linguistic divergences. It is still open to future research whether these differences point to features that could be employed in automatic anaphora resolution. A larger annotated corpus is needed to answer this question in a conclusive way. Models of preferences and divergences in the expression of abstract anaphora are also important for applications such as machine translation.

Another explanation for the observed divergences would be that they are due to idiomatic preferences of the speaker on the one hand and the translator on the other hand and would not be related to different types of anaphors. In a study by [11], it has been investigated to what extent translators differ when they are asked to translate one and the same text. Mismatches that occur between such multiple translations concern syntactic variation, clause subordination vs. anaphorically linked sentences, different argument realization, etc. The same types of mismatches have also been found in comparing original texts and their translations [10]. We think that the Europarl corpus is a suitable database to overcome these objections as it consists of contributions of many speakers and translations by a variety of translators, which is, unfortunately, not documented in the metadata of the corpus.

6 Future Steps

An open question is whether (some of) the cross-linguistic differences can be attributed to differences on the semantic level. In future work, we would like to address the two following hypotheses: (i) English demonstratives conflate different functions of German anaphors, (ii) (Some) differences between both languages could be related to the abstract types of the anaphor and antecedent.

In addition to exploring new features, a larger database will allow us to investigate correlations between already described features such as function, position, and pronoun type. Multivariate analyses could point to hidden preferences and divergences. We expect this deeper approach allows us to explore whether the observed differences can be mapped onto language-specific structures or principles.

Another question not yet investigated is to what extent the use of lexical NP anaphors (e.g., *this situation*) can be exploited to derive features for annotation in a semi-automatic and less subjective way than manual annotation of pronominal anaphors. A further investigation will be on whether the alignments of the parallel corpus can be employed for this endeavor in making use of lexical NP translations to determine the abstract type of a pronominal anaphor in the original text.

References

1. Aduriz, I., Ceberio, K., Díaz, I.D.: Pronominal anaphora in Basque: annotation of a real corpus. In: Proceedings of DAARC-2009. pp. 99–104 (2009)
2. Asher, N.: Reference to Abstract Objects in Discourse. Kluwer Academic Publishers, Boston MA (1993)
3. Azuma, H.: A diachronic view of pronominal reference in English. In: Proceedings of the Second Workshop on Anaphora Resolution (WAR II) (2008)
4. Byron, D.K.: Resolving pronominal reference to abstract entities. In: Proceedings of the ACL-02 conference. pp. 80–87 (2002)
5. Byron, D.K.: Annotation of pronouns and their antecedents: A comparison of two domains (2003), technical report, University of Rochester
6. Dipper, S., Müller, M., Rieger, C., Seiss, M., Zinsmeister, H.: Discourse-deictic anaphora — comparison EN–GE (2011), annotation guidelines
7. Dipper, S., Zinsmeister, H.: Discourse-deictic anaphora (2009), annotation guidelines
8. Dipper, S., Zinsmeister, H.: Towards a standard for annotating abstract anaphora. In: Proceedings of the LREC 2010 Workshop on Language Resources and Language Technology Standards. pp. 54–59. Valletta, Malta (2010)
9. Dipper, S., Zinsmeister, H.: Annotating abstract anaphora. Language Resources and Evaluation (2011), Online First
10. Hedberg, N., Gundel, J.K., Zacharski, R.: Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In: Proceedings of DAARC-2007. pp. 31–36 (2007)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit (2005)

12. Kučová, L., Hajičová, E.: Coreferential Relations in the Prague Dependency Treebank. In: Proceedings of DAARC-2004. pp. 97–102 (2004)
13. Müller, C.: Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In: Proceedings of ACL-07. pp. 816–823 (2007)
14. Navarretta, C.: A contrastive analysis of the use of abstract anaphora. In: Proceedings of DAARC-2007: 6th Discourse Anaphora and Anaphora Resolution Colloquium. pp. 103–109 (2007)
15. Nedoluzhko, A., Mírovský, J., Ocelák, R., Pergler, J.: Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In: Proceedings of DAARC-2009 (2009)
16. Passonneau, R.J.: Getting at discourse referents. In: Proceedings of ACL-89 (1989)
17. Poesio, M., Artstein, R.: Anaphoric annotation in the ARRAU corpus. In: Proceedings of LREC-08 (2008)
18. Pradhan, S., Ramshaw, L., Weischedel, R., MacBride, J., Micciulla, L.: Unrestricted coreference: Identifying entities and events in OntoNotes. In: Proceedings of the IEEE-ICSC (2007)
19. Recasens, M.: Discourse deixis and coreference: Evidence from AnCora. In: Proceedings of the Second Workshop on Anaphora Resolution (WAR II). pp. 73–82 (2008)
20. Schmid, H.: Probabilistic part-of-speech tagging using decision tree. In: Proceedings of International Conference on New Methods in Language Processing (1994)
21. Vieira, R., Salmon-Alt, S., Gasperin, C.: Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In: Proceedings of DAARC-2002 (2002)
22. Webber, B.: A Formal Approach to Discourse. Garland (1979)
23. Webber, B.: Discourse deixis: Reference to discourse segments. In: Proceedings of ACL-88. pp. 113–122 (1988)
24. Webber, B.: Structure and ostention in the interpretation of discourse deixis. Language and Cognitive Processes 6, 107–135 (1991)